Physica A xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Q1 Cross-correlation patterns in social opinion formation with sequential data

Q2 Anindya S. Chakrabarti

Economics Area, Indian Institute of Management, Vastrapur, Ahmedabad 380015, India

ARTICLE INFO

Article history:
Received 5 January 2016
Received in revised form 4 April 2016
Available online xxxx

Keywords:
Online rating
Movie popularity
Social opinion
Sociophysics

ABSTRACT

Recent research on large-scale internet data suggests existence of patterns in the collective behavior of billions of people even though each of them may pursue own activities. In this paper, we interpret online rating activity as a process of forming social opinion about individual items, where people sequentially choose a rating based on the current information set comprising all previous ratings and own preferences. We construct an opinion index from the sequence of ratings and we show that (1) movie-specific opinion converges much slower than an independent and identically distributed (i.i.d.) sequence of ratings, (2) rating sequence for individual movies shows lesser variation compared to an i.i.d. sequence of ratings, (3) the probability density function of the asymptotic opinions has more spread than that defined over opinion arising from i.i.d. sequence of ratings, (4) opinion sequences across movies are correlated with significantly higher and lower correlation compared to opinion constructed from i.i.d. sequence of ratings, creating a bimodal cross-correlation structure. By decomposing the temporal correlation structures from panel data of movie ratings, we show that the social effects are very prominent whereas group effects cannot be differentiated from those of surrogate data and individual effects are quite small. The former explains a large part of extreme positive or negative correlations between sequences of opinions. In general, this method can be applied to any rating data to extract social or group-specific effects in correlation structures. We conclude that in this particular case, social effects are important in opinion formation process.

1. Introduction

In the last decade, there has been a massive change in information sciences and technologies which have enabled the development of interactive platforms facilitating human–computer and human–human interactions. Consumers can share their thoughts and opinions about products they consume. The aggregate information about past choices are useful for online recommendation systems as well as for direct recommendation by extrapolating current opinions expressed online. This is reflected in emergence of a number of services like IMDB, Rotten Tomatoes, Tripadvisor, Yelp, etc. where consumers can directly contribute their product specific personal experiences. Along with verbal descriptions, typically the consumers are also enabled to quantify their opinions by assigning numerical ratings to specific products. These ratings as well as review are made publicly available free of cost which helps the new consumers to make comparisons and choices. For the present paper, we ignore the literary descriptions for analytical convenience and focus on the numerical ratings which at any given time broadly reflect the overall public opinion about a given item or product. This opinion evolves over time and

E-mail address: anindyac@iimahd.ernet.in.

http://dx.doi.org/10.1016/j.physa.2016.06.056
0378-4371/

Please cite this article in press as: A.S. Chakrabarti, Cross-correlation patterns in social opinion formation with sequential data, Physica A (2016), http://dx.doi.org/10.1016/j.physa.2016.06.056

the evolution is directly influenced by the perceived quality of the product as well as various social factors including word-of-mouth effects [1]. In this paper, we aim to understand the temporal evolution of the mutual interaction between various products within a given product category. In particular, we focus on movies because of availability of large scale rating data. We show that the opinions about movies evolve substantially over time and reach movie-specific equilibrium values which are significantly different from those arising from arbitrary rating schemes. The rate of convergence is also much slower. Next, we study the evolution of opinion across movies and we present evidence that there is substantial bimodality in the correlation structure of movies indicating that the evolution of opinions about movies is either highly correlated or highly anti-correlated. Our analysis shows that the social effects are very pronounced in both modes.

This finding complements the existing work on opinion formation. Ref. [2] presented evidence of bimodality in income distribution in cross-sectional data of movie income. This clearly reflects a certain polarization in success of products in terms of measurable quantities like income (see also Refs. [3,4]). Ref. [5] shows that the bimodal pattern in the income distribution (both opening and lifetime) is robust over significantly long time span and presented an agent-based model to explain the origin of such bimodality. Such a divergence has also been found in electoral data [6]. However, these may not necessarily reflect a divergence in social opinion being the cause. This point requires some clarifications. A standard theory of diverging success builds on the idea of word-of-mouth learning which is essentially a demand-side story. Given significant amount of feedback effect, the demand-side i.e. the consumers might treat products very differently creating a divergence in the eventual success of the product. However, as Ref. [5] argues, the divergence may in fact arise from supply-side competition i.e. the race between products to occupy niche in the competitive landscape of the market. Thus both explanations are feasible.

Detailed studies have shown existence of diverse social patterns from large-scale data including but not limited to, voting behavior [7], sexual contacts [8] and citation networks [9]. Parallel developments have been made in theoretical understanding as well (see e.g. Refs. [10–13,1,14,15]). See Ref. [16] for a detailed review of theoretical development in this direction.

2. Cross-correlation structure

First we describe construction of an index of movie opinion. Consider a panel data of rating over movies $\{r_{it}\}_{i\in N, t\in T}$ where N is the number of movies considered and T is the length of the rating time-series i.e. the number of ratings received. Suppose the information set of the tth rater rating the ith movie is denoted by a vector x_{it}. Then the rating for the ith movie at the tth time-point is a mapping from the set of characteristics to the set of possible ratings, $R = \{0.5, \ldots, 5\}$. This can be denoted by some function f that the tth rater uses the rate of the ith movie,

$$f_{it} : x_{it} \to R. \tag{1}$$

Note that each rating refers to individual beliefs and assessments about a movie by one single individual. Thus we need to aggregate the information to have an idea about the evolution of the common 'opinion' about a movie. We are ignoring here the fact that such ratings may not reflect the actual quality of a movie as the actual quality is a very slippery concept. We construct the following indicator of aggregate social opinion at time t,

$$S_{it} = \frac{1}{t}\sum_{t'\le t} r_{it'}. \tag{2}$$

Thus the social opinion is just an expanding average over the movie ratings $r_{it'}$ across time. It is worthwhile to see if this converges at all and also see how it compares with an arbitrary rating. The first point can be dealt with mathematically. Suppose all individuals follow the same algorithm f to rate movies by considering their own opinion and the past history of ratings (which also includes the possibility of considering the probability density function of past ratings). The decision can be purely stochastic as well.

Note that the each rating is a random variable and hence so is the social opinion (from Eq. (2)). Thus to show convergence of social opinion, we need to show that the right hand side of Eq. (2) converges. One can, in principle, apply the usual *law of large numbers*(LLN) for a random i.i.d. sequence of ratings, for which the convergence is evident. For example, the Chebychev's LLN requires finite mean and variance of r_{it}. Kolmogorov's version requires finite mean of both r_{it} and absolute value of it. All such conditions are satisfied by the rating data under study. Here the problem is that the rating sequence could well be dependent and it is actually very likely that while assigning rating to a movie individuals are influenced by the prevailing social opinion about the opinion. To deal with this case, we note that the range of rating is finite and bounded. This implies that the rating r is integrable random variable ($r_i \in L_1$) with finite mean $E(r_i)$. By making an assumption that $E(r_{it}) = E(r_i)$, one can apply the strong law of large numbers for dependent sequence (dependent but uncorrelated or even with further weakening, serial correlation) to S_{it} and still show convergence to $E(r_i)$. See Theorem 22 in Ref. [17] for a strong version of the theorem with the necessary conditions being stationarity and ergodicity. Thus the convergence issue can be settled. Note that movie-specific opinion can converge to different values which is also seen in Fig. 1.

Fig. 1 provides visual description of the data. The top-left panel shows evolution of three opinion sequences for three movies. The top-middle panel shows three opinion sequences generated from random ratings. It is evident that actual opinion fluctuates much more than synthetic data and hence, the convergence is also much slower. It is worthwhile to

Please cite this article in press as: A.S. Chakrabarti, Cross-correlation patterns in social opinion formation with sequential data, Physica A (2016), http://dx.doi.org/10.1016/j.physa.2016.06.056